RESEARCH ARTICLE

JENER Journal of Empirical and Non-Empirical Research
Volume 2, Issue 3, Page 447-453 | Article No. 63
DOI: 10.4898/jener.v2i3.a63
ISSN: 7895-9216
Journal homepage: https://jenerjournal.com
Journal email: editor@jenerjournal.com

JENER Journal
JENER Journal of Empirical and Non-Empirical Research

Self-supervised Pre-training for Histological Image Transformer

Ok Chol Ri[1], Song Il An[1] & Ho Kim[1]

[1]Faculty of Artificial Intelligence, Kim Il Sung University, Pyongyang, Democratic People's Republic of Korea

ARTICLE INFORMATION

Article history:
Published: March 2026

Keywords:
Histological image
Self-supervised learning
Deep learning

ABSTRACT

Image Transformer has recently achieved significant progress for natural image understanding, either using supervised (ViT, DeiT, etc.) or self-supervised (BEiT, MAE, etc.) pre-training techniques. In this paper, we propose HiT, a self-supervised pre-trained Histological Image Transformer model using large-scale unlabeled histological images for medical image processing tasks, which is essential since no supervised counterparts ever exist due to the lack of human-labeled histological images. We leverage HiT as the backbone network in a variety of vision-based histological image processing tasks. Experiment results have illustrated that the self-supervised pre-trained HiT model the new state-of the-art results on these downstream tasks, e.g. histological image classification on SIPaKMeD database achieved an accuracy of 97.45% and 99.29% for 5-class and 2-class classifications, respectively.

1. Introduction

Recently, deep convolutional network (CN), an instance of Deep Learning (DL) architectures, has shown that it is superior to the other non-deep learning based approaches in image analysis. DL is the data-driven and end-to-end learning approach which learns high-level structure features from only the pixel intensities that are useful for differentiating objects by a classifier. Recently, it has been successfully employed for medical image analysis with various applications [1–9].

The approaches based on DL have been evoked great interests in the histological image analysis community since the pioneer made a research in [10]. Histopathology represented an excellent use for application of deep learning strategies when its size and complexity were given. In [10], a popular sixlayer CN, which is also called “ConvNet” is employed for mitotic detection. The model was a patch- wise training process for pixel-wise labeling. It was first trained with a great amount of context patches. Basically, there were two types of context patches: foreground patches whose central pixels are located within target objects and background patches whose central pixels are located around the neighborhood pixel of the target objects. After training, the model was employed to predict the central pixel of chosen patches being targeted objects or not. Recently, much effort has been focused on nuclear detection or segmentation [11–13]. In terms of nuclear detection, a Stacked Sparse Auto encoder(SSAE)-based model was employed in [14] for discriminating nuclear and non-nuclear patches. Then integrating with the sliding window operation the SSAE model was further utilized for automated nuclear detection from high-resolution histological images in [11].

In [12], a Spatially Constrained CN was presented to nucleus detection. This segmentation free strategy can detect and classify different nuclear types simultaneously on colorectal adenocarcinoma images. The CN involves convolutional and subsampling operations to learn a set of locally connected neurons through local receptive fields for feature extraction. Therefore, CN is good for capturing contextual information. Based on this contexture information, a pixel-wise based CN was developed for pixel-wise segmentation of nuclear regions in [13]. Pixel-wise segmentation is different from patch-wise classification since pixel-wise segmentation aims at predicting class label of each pixel in an image based on a local patch around that pixel [10], while patch-wise classification aims for assigning a single label to the entire image patch [15]. Therefore, pixel-wise classification is more challenged. In [16], the authors employed the convolutional auto-encoder neural network architecture with auto-encoder for histopathological image representation learning. Then, the softmax classification is employed for classifying regions of cancer and non-cancer.

But DL-based approaches have some challenges for medical image processing especially for histological image.

DL needs considerable training data because the data set’s size and quality significantly impact the classifier’s effectiveness. But a lack of data is one of the biggest obstacles for using DL in medical imaging. Generating significant amounts of medical imaging data is challenging because eliminating human error takes a great deal of work for experts. Large medical imaging data sets are difficult to construct because annotating the data takes a great deal of time and effort from a single expert to many experts to eliminate human error. The absence of substantial training datasets has made it challenging to construct deep-learning models for medical imaging, which was the first problem we saw in our studies. Most reviewed studies evaluated and assessed these using various datasets that are collected from the cancer research organizations or clinics privately. The main issue with this method is that it is impossible to compare how well such models functioned in several investigations.

The absence of benchmarks provided a hurdle and highlighted a lack of flexibility.

Another issue with specific papers is using data expansion techniques rather than transferring learning to minimize overfitting.

Techniques for breast cancer categorization using unsupervised grouping: The supervised learning method was used to classify breast cancer in most of the selected primary papers. These strategies have provided superior results when labeled images are used

throughout the training. However, finding breast cancer images with precise, medically labeled criteria might be difficult. There are frequently many unidentified medical images available. Despite this is useful knowledge sources, many blank labels cannot be used for supervised learning. Therefore, there is a pressing need for a cancer categorization model that may be created using several grouping techniques without supervision.

Methodology of reinforcement learning for breast cancer classification: The fundamental issue is a lack of sufficient breast cancer image examples to depict all types of breast cancer. Creating a machine learning model that simultaneously learns from its surroundings can be difficult. Therefore, systems for identifying breast cancer from medical photos can perform and be more effective when employing a learning-based reinforcement model.

Reliability of data collection techniques: The robustness issue of various clinical and technical circumstances must be addressed to integrate new datasets gradually. Different image acquisition scanners, lighting configurations, sizes, and views across many picture modalities, and varying presentation aspects of the coloring and enlargement factors, are a few examples of these modifications.

Despite its significance in medical picture segmentation, the segmentation's influence still falls short of what is required for practical use.

Experiment results have illustrated that the pre-trained HiT model has outperformed the existing supervised and self-supervised pre-trained models and achieved new state-of-the-art on these tasks. The contributions of this paper are summarized as follows:

(1) We generate large number of unlabeled histological images based on StyleGAN for self- supervised learning to resolve weak datasets problem.

(2) We propose HiT, a self-supervised pre-trained histological image Transformer model, which can leverage large-scale unlabeled histological images for pre-training.

(3) We leverage the pre-trained HiT models as the backbone for a histological image classification task on SIPaKMeD database.

2. Related Work

Image Transformer has recently achieved significant progress in computer vision problems, including classification, object detection, and segmentation. In [17], it was firstly applied the standard Transformer directly to images with the fewest modifications. They split an image into 16 × 16 patches and provide the sequence of linear embedding of these patches as an input to a Transformer named ViT. The ViT model is trained on image classification in a supervised fashion and outperforms the ResNet baselines. [18] proposed data-efficient image transformers & distillation through attention, namely DeiT, which solely relies on the ImageNet dataset for supervised pre-training and achieves SOTA results compared with ViT. [19] proposed a hierarchical Transformer whose representation is computed with shifted windows. The shifted windowing scheme brings efficiency by limiting self- attention computation to non-overlapping local windows while also allowing for cross-window connection. In addition to supervised pre-trained models, [20] trained a sequence Transformer called iGPT to auto-regressively predict pixels without incorporating knowledge of the 2D input structure, which is the first attempt at self-supervised image transformer pre-training.

After that, self-supervised pre-training for image Transformer became a hot topic in computer vision. [21] proposed DINO, which pre-trains the image Transformer using self-distillation with no labels. [22] proposed MoCov3 that is based on Siamese networks for self-supervised learning. More recently, [23] adopted a BERT-style pre-training strategy, which first tokenizes the original image into visual tokens, then randomly masks some image patches and feeds them into the backbone Transformer. Similar to the masked language modeling, they proposed a masked image modeling task as the pre-training objective that achieves SOTA performance. [24] presented a self-supervised framework iBOT that can perform masked prediction with an online tokenizer. The online tokenizer is jointly learnable with the MIM objective and dispenses with a multi-stage pipeline where the tokenizer is pre-trained beforehand. Due to the lack of large-scale human-labeled datasets in histological image processing domain, existing approaches are usually based on the ConvNets models that are pre-trained with ImageNet/COCO datasets. Then, the models are continuously trained with task-specific labeled samples. To the best of our knowledge, the pre-trained HiT model is the first large-scale self- supervised pre-trained model for histological image processing tasks. Meanwhile, it can be further leveraged for the multimodal pre-training for medical image processing.

3. Histological Image Transformer

In this section, firstly, we present the generating method of unlabeled histological images, the architecture of HiT and the pre-training procedure. Then, we describe the application of HiT models in different downstream tasks.

3.1 Generating Unlabeled Histological Images

The majority of studies on histopathology image analysis, according to the literature, are based on small datasets that are often not shared with the scientific community.

For example, Break His dataset is introduced in this. At four distinct magnifications (40×, 100×, 200×, and 400×), 82 patients provided 7909 microscopic photos of breast tumor tissue that were clinically realistic. These images were collected for BreaKHis. It now has 2480 benign samples and 5429 cancerous ones. All information was made anonymous. Hematoxylin and eosin (HE)-stained breast tissue biopsy slides were used to create the samples. Pathologists from the P&D Lab obtained the samples through surgical (open) biopsy (SOB), prepared them for histological analysis, and labeled them. From this point we need to generate lots of histological images for self-supervised pretraining process. We introduced StyleGAN[25], developed by NVIDIA AI Lab, for generating unlabeled histological images.

We used StyleGAN3 config-r and Break His dataset(including 7909 photos) to generate unlabeled images. Real and generated histological images are shown in Figure 1 and Figure 2.

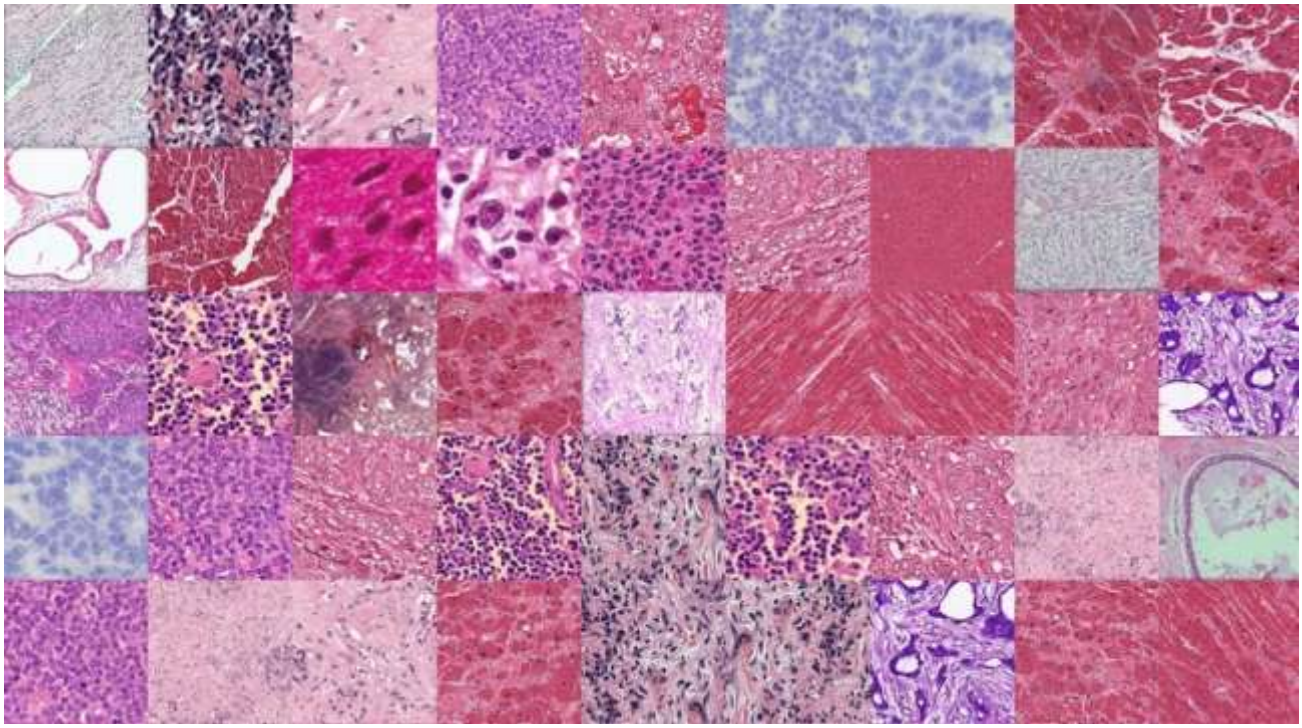

Figure 1: Real Histological Images

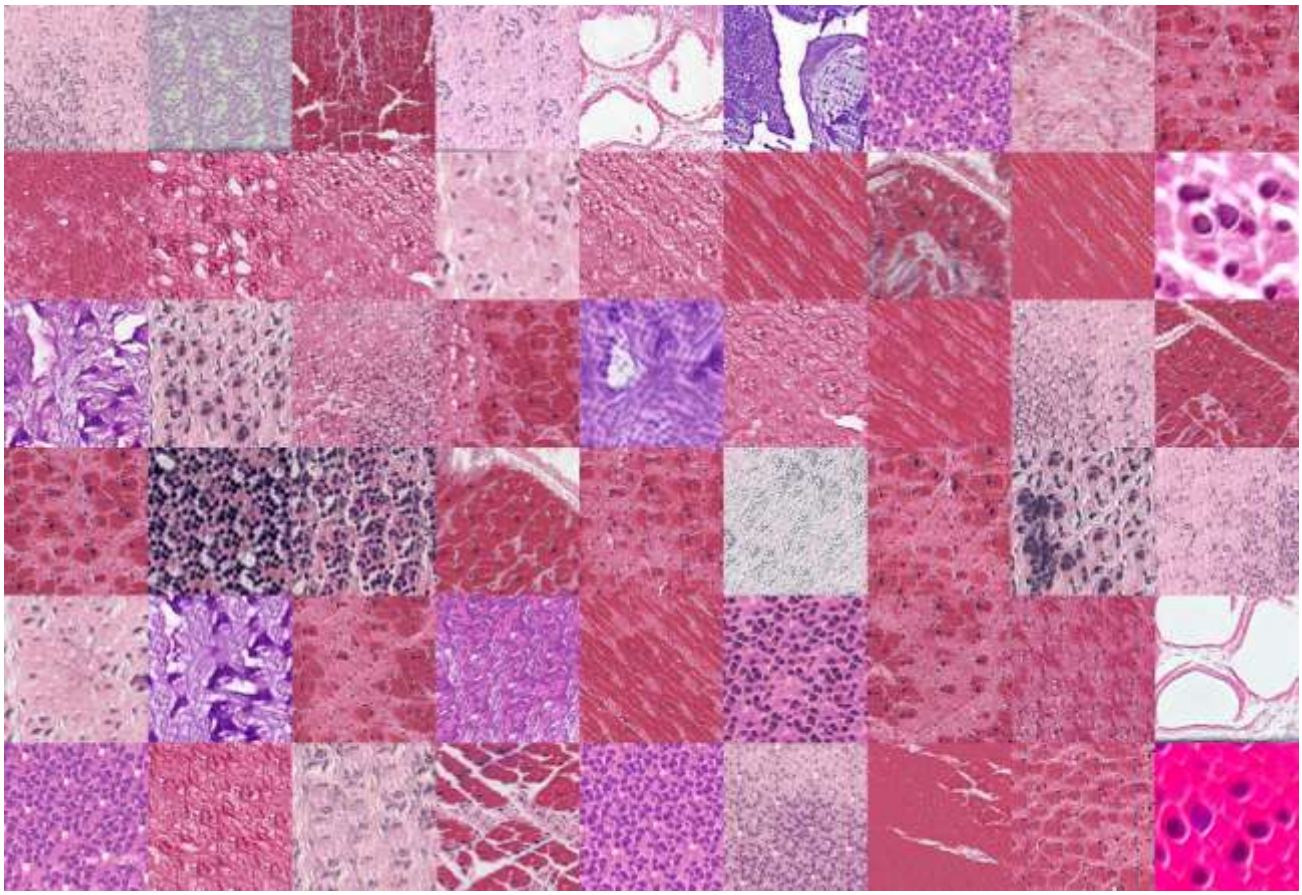

Figure 2: Generated Histological Images by StyleGAN

3.2 HiT Model Architecture

Following ViT [17], we use the vanilla Transformer architecture [26] as the backbone of HiT. We divide a histological image into non-overlapping patches and obtain a sequence of patch embedding.

After adding the 1d position embedding, these image patches are passed into a stack of Transformer blocks with multi-head attention. Finally, we take the output of the Transformer encoder as the representation of image patches, which is shown in Figure 3.

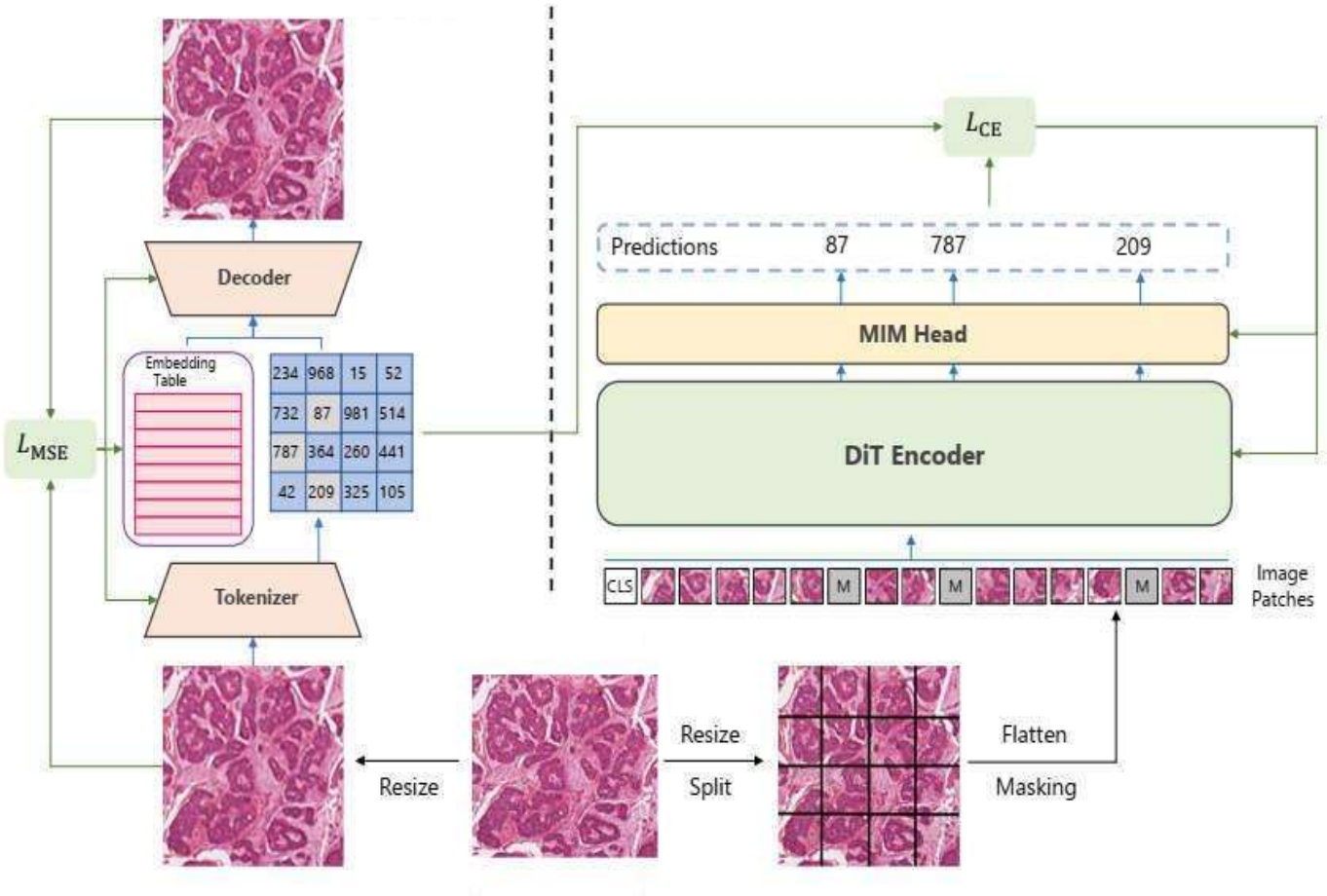

Figure 3: The model architecture of HiT with MIM pre-training

3.3 Pre-training

Inspired by BEiT [23], we use Masked Image Modeling (MIM) as our pre-training objective. In this procedure, the images are represented as image patches and visual tokens in two views respectively.

During pre-training, HiT accepts the image patches as input and predicts the visual tokens with the output representation. Like text tokens in natural language, an image can be represented as a sequence of discrete tokens obtained by an image tokenizer. BEiT uses the discrete variational auto-encoder (dVAE) from DALLE [27] as the image tokenizer, which is trained on a large data collection including 400 million images. However, there exists a domain mismatch between natural images and histological images, which makes the DALL-E tokenizer not appropriate for the histological images.

Therefore, to get better discrete visual tokens for the histological image domain, we train a dVAE on our generated dataset from stylegan that includes 42 million histological images.

To effectively pre-train the HiT model, we randomly mask a subset of inputs with a special token [MASK] given a sequence of image patches. The HiT encoder embeds the masked patch sequence by a linear projection with added positional embedding, and then contextualizes it with a stack of Transformer blocks. The model is required to predict the index of visual tokens with the output from masked positions. Instead of predicting the raw pixels, the masked image modeling task requires the model to predict the discrete visual tokens obtained by the image tokenizer.

3.4 Fine-tuning

We fine-tune our model on publicly accessible SIPaKMeD database containing 966 WSI pap smear images and 4,049 images of handcrafted cropped cells [28]. An optical magnifying device (OLYMPUS BX53F) with a camera having a charge-coupled device (CCD) sensor (Lumenera's INFINITY-1) has been used to capture these pictures. The dataset is categorized into 5 classes by clinical professionals. The classes "superficial-intermediate (SI)" and "parabasal (P)" refer to "normal," images sorted as "koilocytotic (K)" and "dyskeratotic (D)" indicate "abnormal," and the remaining "metaplastic (M)" belongs to have "benign" cells. The experiment is performed on WSI slides and grouped into 5 class and 2 class (normal and abnormal).

Furthermore, the proposed framework is evaluated using liquid-based cytology (LBC) data available online at Mendeley data [29]. Based on the Bethesda system, the collection includes 963 WSI LBC high-resolution images organized into four sets of classes: "no squamous intraepithelial lesion (NILM)," "low-grade squamous intraepithelial lesion (LSIL)," "high-grade squamous intraepithelial lesion (HSIL)," and "squamous cell carcinoma (SCC)." The "NILM" indicates a "normal" grade, while the "LSIL," "HSIL," and "SCC" refer to "abnormal."

For cancer detection based on image classification, we use average pooling to aggregate the representation of image patches. Next, we pass the global representation into a simple linear classifier.

4. Experiments

4.1 Settings

Pre-training Setup. We pre-train HiT on the IIT-CDIP Test Collection 1.0 [30]. We pre-process the dataset by splitting multi-page documents into single pages, and obtain 42 million histological images. We also introduce random resized cropping to augment training data during training. We train our HiT-B model with the same architecture as the ViT base: a 12-layer Transformer with 768 hidden sizes, and 12 attention heads. The intermediate size of feed-forward networks is 3,072. A larger version, HiT-L, is also trained with 24 layers, 1,024 hidden sizes, and 16 attention heads. The intermediate size of feed-forward networks is 4,096.

The dVAE Tokenizer. BEiT borrows the image tokenizer trained by DALL-E, which is not aligned with the histological image data. In this case, we fully utilize the 42 million histological images in the IIT-CDIP dataset and train a document dVAE image tokenizer to obtain the visual tokens. Like the DALL-E image tokenizer, the histological image tokenizer has the codebook dimensionality of 8,192 and the image encoder with three layers. Each layer consists of a 2D convolution with a stride of 2 and a ResNet block. Therefore, the tokenizer eventually has a down sampling factor of 8.

In this case, given a 112×112 image, it ends up with a 14×14 discrete token map aligning with the 14×14 input patches.

We implement our dVAE codebase from open-sourced DALL-E implementation1 and train the dVAE model with the entire IITCDIP dataset containing 42 million histological images. The new dVAE tokenizer is trained with a combination of a MSE loss to re-construct the input image, and a perplexity loss to increase the use of the quantized codebook representations. The input image size is 224×224, and we train the tokenizer with a learning rate of 5e-4 and a minimum temperature of 1e-10 for 3 epochs.

Fine-tuning on SIPaKMeD. We evaluate the pre-trained HiT models and other image backbones on SIPaKMeD for histological image classification. We fine-tune the image transformers for 100 epochs with a batch size of 128 and a learning rate of 1e-3. For all settings, we resize the original images to 224 × 224 with the Random Resized Crop operation.

4.2 Results

Table 1 depicts the results obtained from different classifiers and our proposed model on SIPaKMeD WSI data for classification into 5 class and 2 class.

Table 1: Performance metrics for different fine-tuned classifiers and proposed model on SIPaKMeD WSI data

Data	Models	Accuracy (%)	Precision (%)	Recall (%)	F-score (%)
SIPaKMeD WSI 5-class	VGG-16	94.89	95.88	95.77	95.83
	ResNet-152	93.37	93.76	94.66	94.21
	DenseNet-169	90.82	91.94	92.06	92
	Proposed model	97.45	97.94	98.08	98.01
SIPaKMeD WSI 2-class	VGG-16	97.16	95.7	100	97.8
	ResNet-152	96.45	94.62	100	97.24
	DenseNet-169	94.33	93.55	97.75	95.6
	Proposed model	99.29	98.92	100	99.46

Here is the performance of the individual fine-tuned models of VGG-16, ResNet-152, and DenseNet-169 are confronted with our proposed model based on HiT pre-training. Our proposed model achieves better results than the other classifiers with an accuracy of 97.45% and 99.29% for 5-class and 2-class classifications, respectively. For 5-class classification, our model predicts the papsmear images with a precision score of 97.94%, a recall value of 98.08%, and an F-score of 98.01%.

This model outperforms the other classifiers with the precision, recall, and F-score values of 98.92%, 100%, and 99.46%, respectively, in 2 class classification.

The accuracy and loss curve for SIPaKMED 5-class classification of our proposed model based on HiT pre-training are shown in Figure 4.

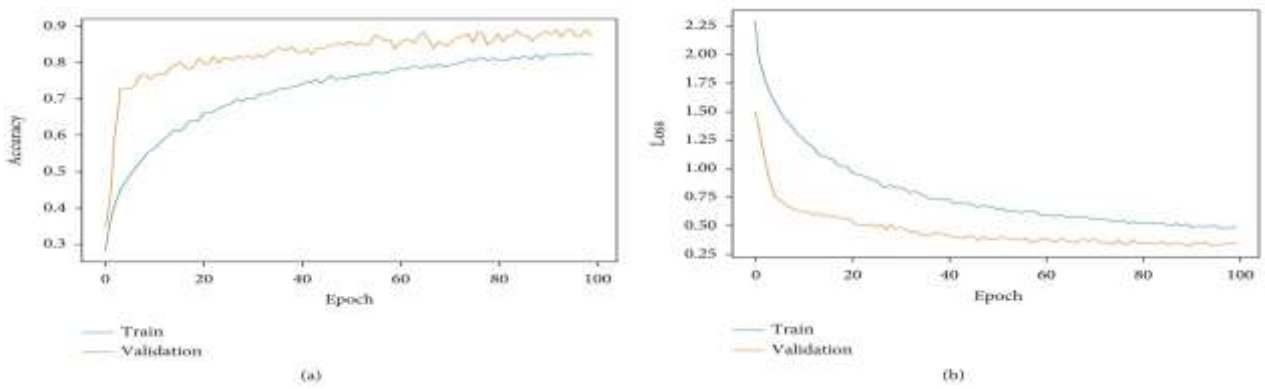

Figure 4: Fine-tuned classifier
(a) Accuracy curve, (b) Loss curve for SIPaKMED 5-class classification

5. Conclusion

In this paper, we present HiT, a self-supervised foundation model for general histological image processing tasks. The HiT model is pre-trained with largescale unlabeled histological images that cover a variety of templates and formats, which is ideal for downstream histological image processing tasks. We evaluate the pre-trained HiT on histological image classification task. The proposed model based on HiT pre-training is evaluated on SIPaKMeD data and gives an accuracy of 97.45% for 5-class classification and 99.29% for 2-class classification. Moreover, the experiment performed on LBC WSI data provides 99.49% accuracy. The precise recognition of infected WSI images enables experts to perform a more in-depth analysis of cells within the images. The futuristic approach to this method involves the utilization of more optimal feature selection algorithms, progressive resizing, and advanced ensemble methods to further improve model performance and computation cost-cutting.

References

[1] Anthimopoulos M, Christodoulidis S, Ebner L, Christe A, Mougiakakou S. 2016. Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. IEEE Trans Med Imaging 35(5):1207–1216

[2] Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, Summers RM. 2016. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 35(5):1285–1298

[3] Albarqouni S, Baur C, Achilles F, Belagiannis V, Demirci S, Navab N. 2016. AggNet: deep learning from crowds for mitosis detection in breast cancer histology images. IEEE Trans Med Imaging 35(5):1313–1321.

[4] Pereira S, Pinto A, Alves V, Silva CA. 2016. Brain tumor segmentation using convolutional neural networks in MRI images. IEEE Trans Med Imaging 35(5):1240–1251.

[5] Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, Liang J. 2016. Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans Med Imaging 35(5):1299–1312.

[6] Moeskops P, Viergever MA, Mendrik AM, de Vries LS, Benders MJNL, Isgum I. 2016. Automatic segmentation of MR brain images with a convolutional neural network. IEEE Trans Med Imaging 35(5):1252–1261.

[7] Setio AAA, Ciompi F, Litjens G, Gerke P, Jacobs C, van Riel SJ, Wille MMW, Naqibullah M, Snchez CI, van Ginneken B. 2016. Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks. IEEE Trans Med Imaging 35(5):1160–1169.

[8] Van Grinsven MJJP, van Ginneken B, Hoyng CB, Theelen T, Snchez CI. 2016. Fast convolutional neural network training using selective data sampling: application to hemorrhage detection in color fundus images. IEEE Trans Med Imaging 35(5):1273–1284.

[9] Liskowski P, Krawiec K. 2016. Segmenting retinal blood vessels with deep neural networks. IEEE Trans Med Imaging PP(99):1–1.

[10] Ciresan DC et al. 2013. Mitosis detection in breast cancer histology images with deep neural networks. In: MICCAI 2013. LNCS, vol 8150. Springer, Berlin, pp 411–418.

[11] Xu J, Xiang L, Liu Q, Gilmore H, Wu J, Tang J, Madabhushi A. 2016. Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images. IEEE Trans Med Imaging 35(1):119–130.

[12] Sirinukunwattana K, Raza SEA, Tsang YW, Snead D, Cree IA, Rajpoot NM. 2016. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. IEEE Trans Med Imaging 35:1196.

[13] Xing F, Xie Y, Yang L. 2016. An automatic learning-based framework for robust nucleus segmentation. IEEE Trans Med Imaging 35(2):550–566.

[14] Xu J, Xiang L, Hang R, Wu J. 2014. Stacked sparse autoencoder (SSAE) based framework for nuclei patch classification on breast cancer histopathology. In: ISBI.

[15] Xu J, Luo X, Wang G, Gilmore H, Madabhushi A. 2016. A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images. Neurocomputing 191:214–223.

[16] Cruz-Roa A et al. 2013. A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection. In: MICCAI 2013, vol 8150. Springer, Berlin, pp 403–410.

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR (2021)

[18] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In International Conference on Machine Learning. PMLR, 10347–10357

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. arXiv:2103.14030 [cs.CV]

[20] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. 2020. Generative Pretraining From Pixels. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event (Proceedings of Machine Learning Research, Vol. 119). PMLR, 1691–1703.

[21] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging Properties in Self-Supervised Vision Transformers. arXiv:2104.14294 [cs.CV]

[22] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. FUNSD: A Dataset for Form Understanding in Noisy Scanned Documents. 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW) 2 (2019), 1–6.

[23] Hangbo Bao, Li Dong, and Furu Wei. 2021. BEiT: BERT Pre-Training of Image Transformers. arXiv:2106.08254 [cs.CV]
[24] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. 2021. iBOT:Image BERT Pre-Training with Online Tokenizer. arXiv:2111.07832 [cs.CV]
[25] Tero Karras, Samuli Laine, Miika Aittala and Janne Hellsten. 2020. Analyzing and Improving the Image Quality of StyleGAN. arXiv:1912.04958v2 [cs.CV]
[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008.
[27] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. arXiv:2102.12092 [cs.CV]
[28] M. E. Plissiti, P. Dimitrakopoulos, G. Sfikas, C. Nikou, O. Krikoni, and A. Charchanti, "SIPAKMED: a new dataset for feature and image based classification of normal and pathological cervical cells in pap smear images," in 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 3144–3148,Athens, Greece, 2018.
[29] E. Hussain, L. B. Mahanta, H. Borah, and C. R. Das. 2020. "Liquid based-cytology pap smear dataset for automated multi-class diagnosis of pre-cancerous and cervical cancer lesions," Data in Brief, vol. 30.
[30] D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, and J. Heard. 2006. Building a Test Collection for Complex Document Information Processing. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Seattle, Washington, USA) (SIGIR '06). ACM, New York, NY, USA, 665–666. https://doi.org/10.1145/1148170.1148307.
[31] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6980.
[32] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger. 2016. Deep Networks with Stochastic Depth. In ECCV.